

# Replication and Automation of Expert Judgments: Information Engineering in Legal E-Discovery

Bruce Hedin  
H5  
San Francisco, CA 94105, USA  
bhedin@h5.com

Douglas W. Oard  
College of Information Studies &  
UMIACS CLIP Lab  
University of Maryland  
College Park, MD 20742, USA  
oard@umd.edu

**Abstract**—The retrieval of digital evidence responsive to discovery requests in civil litigation, known in the United States as “e-discovery,” presents several important and understudied conditions and challenges. Among the most important of these are (i) that the definition of responsiveness that governs the search effort can be learned and made explicit through effective interaction with the responding party, (ii) that the governing definition of responsiveness is generally complex, deriving both from considerations of subject-matter relevance and from considerations of litigation strategy, and (iii) that the result of the search effort is a set (rather than a ranked list) of documents, and sometimes a quite large set, that is turned over to the requesting party and that the responding party certifies to be an accurate and complete response to the request. This paper describes the design of an “Interactive Task” for the Text Retrieval Conference’s Legal Track that had the evaluation of the effectiveness of e-discovery applications at the “responsive review” task as its goal. Notable features of the 2008 Interactive Task were high-fidelity human-system task modeling, authority control for the definition of “responsiveness,” and relatively deep sampling for estimation of type 1 and type 2 errors (expressed as “precision” and “recall”). The paper presents a critical assessment of the strengths and weaknesses of the evaluation design from the perspectives of reliability, reusability, and cost-benefit tradeoffs.

**Index Terms**—Human-machine cooperation and systems, Information retrieval, Search methods, User modeling, Legal factors

## I. INTRODUCTION

The increasingly large volumes of Electronically Stored Information (“ESI”) that must be produced by parties to civil litigation pose challenges to present search practices and Information Retrieval (IR) technology. In response to this need, the industry has been investigating a range of different tools and methods that may help litigants meet these challenges. What those seeking to use these tools and methods would like is a way to evaluate their effectiveness, and meeting that need calls for an evaluation design that accurately models the conditions and objectives litigants encounter in the real world.

The “production” (i.e., the provision) of digital evidence in civil litigation, known in the United States as “e-discovery,” is characterized by a number of distinctive conditions. Among the most important of these are (i) that the definition of responsiveness that governs the search effort can be learned and made explicit through effective interaction with the responding party, (ii) that the governing definition of responsiveness is generally

complex, deriving both from considerations of subject-matter relevance and from considerations of litigation strategy, (iii) that the result of the search effort is a set (rather than a ranked list) of documents that the responding party certifies to be an accurate and complete response to the request, and (iv) that, in some cases, these sets can be quite large. In this paper, we examine a method for evaluating the effectiveness of information-retrieval methods when applied in the conditions that characterize the task of e-discovery.

The paper proceeds as follows. We begin (Section II) with a characterization of the salient objectives and conditions of the specific domain that is the focus of our study, e-discovery for litigation or regulatory compliance. We then in Section III review a recently developed protocol for evaluating the effectiveness of IR systems applied in this domain. Section IV then reports on our experience with applying this protocol in practice. Finally, Section V draws on that experience to reflect on the strengths and weaknesses of the approach that we have described.

## II. ON OBJECTIVES AND CONDITIONS IN E-DISCOVERY

There are a number of uses to which an attorney may put an IR system over the course of a lawsuit or investigation. The attorney may need to conduct a search of case law in order to find prior rulings and opinions salient to the matter being litigated. The attorney may wish to carry out exploratory probes of a client’s collected documents in order to test initial hypotheses and build a theory of the case. The attorney may need to review a client’s collected documents in order to identify those responsive to requests for production from the opposing party. The attorney will likely want to conduct more narrowly defined searches within the sets of responsive documents (both those of the client and those of the opposing party) in order to prepare for depositions or to fill in gaps in an otherwise well-developed story. And these are just a few. For purposes of this paper, we focus on one of these tasks, the one that generally entails the greatest cost and risk for the client (at least in terms of sanctions from the court): the review of documents for responsiveness to a request for production. For a general discussion of the challenges increasing volumes of ESI pose for the legal profession and of some best practices in applying search and retrieval methods in meeting these

challenges, see The Sedona Conference's commentary on the topic [1].

#### *A. Objectives of Responsive Review*

When a party is served with a request for the production of documents, typically an itemized list of individual requests for documents of certain types or on certain topics, the party will respond to the request with certain objectives in mind. The responding party is acting, in one sense, as an intermediary for the requesting party, endeavoring to locate the documents that are responsive to the latter's requests. In this instance, however, the intermediary (the responding party), is also guided by specific objectives of its own, relating, for example, to questions of the defensibility of the search process employed and to the identification, and withholding, of privileged material.

The responding party will be guided, in part, by accuracy objectives. When served with a discovery request, the party is under an obligation, unless grounds can be cited either for not responding or for only partially responding to the request, to respond in a manner that is complete and accurate, commensurate with a reasonable good-faith effort. What this means, in terms of the accuracy objectives of a responsive review, is that both high recall (i.e., few type 2 errors) and high precision (i.e., few type 1 errors) matter. The party must make an effort both to avoid missing any documents genuinely responsive to a given request and to avoid producing documents that are not genuinely responsive to the request.

The party will also be guided by pragmatic objectives. Almost any document request will leave some scope for interpretation, and how broadly or narrowly a party interprets a request may be influenced by considerations other than simple considerations of relevance. A party may take a broad view of responsiveness, producing documents it may well believe are not genuinely relevant to a request, simply in order to avoid being challenged for underproduction. A party may decide to take a narrow view of responsiveness, constraining the production to documents it believes are genuinely relevant, in order to minimize the risk of disclosing potentially-damaging documents it arguably could have held back.

If, therefore, a text retrieval methodology is to help a party meet its obligations and objectives in responding to a document request, it must, on the one hand, be capable of achieving simultaneous high recall and high precision and, on the other, be susceptible of being steered toward the specific pragmatic objectives that prevail in a given circumstance.

#### *B. Conditions of Responsive Review*

As noted above, document requests are typically underspecified, leaving considerable scope for differences in interpretation. While, however, a large number of interpretations of responsiveness will generally be consistent with a given document request, and while the scope of one interpretation may vary widely from the scope of another, there is, in terms of executing the review and retrieval task, only one interpretation that matters, that of the party to the litigation (or, more typically, that of the attorney representing the party to the litigation).

The responsibility for weighing all interpretations and arriving at a coherent conception of responsiveness belongs typically to the senior litigator, who, in representing the party to the litigation, bears ultimate responsibility for quality of the document production and who must certify to the court that it is, commensurate with a reasonable good-faith effort, accurate and complete. In some cases, the authority for relevance may in fact turn out to be more than one individual; a litigation team may allocate responsibilities among its members in any of a number of ways and in some instances that may mean that authority for relevance determinations is dispersed among a group of individuals rather than concentrated in a single attorney. Such dispersal does not change the fundamental condition, however, for the dispersal of authority extends, at most, to a small, coordinated, team of individuals whose views on relevance can be canvassed, documented, and reconciled in clarifying the intent and scope of a request.

By way of contrast, it is worth noting that this is quite different from the situation with Web search, for example, in which a search engine typically has little or no access to information about what the user actually wishes to find other than the queries that they type and perhaps some of the links that they follow from their result sets. In this circumstance, although each user of a Web search engine is a single authority for what is relevant to their actual information need, the search engine must strive to serve all of those users equally well, and to do so without much evidence. The situation in e-discovery is quite different because authoritative descriptions of, and assessments of, relevance can actually be made available to the provider of review services.

What, then, in a responsive review, the provider of retrieval services is asked to do is to replicate, across the full set of collected documents, a single conception of relevance, the conception that the senior litigator (or litigation team) has concluded is best suited to meeting their client's accuracy obligations and pragmatic objectives.

A second fact that conditions the way in which a responsive review can be conducted is that the conception of relevance may gain in clarity and even evolve over the course of the review. At the outset of the retrieval effort, the attorney who is the authoritative source for relevance criteria cannot have perfect knowledge of all of the subject matter pertinent to all of the document requests nor can the attorney have perfect knowledge of all the ways the pertinent subject matter will manifest itself in the documents being searched. In this initial, imperfect, state of knowledge, the attorney will generally be able to give guidance as to what is relevant, but that guidance may be somewhat provisional and be of limited specificity. In the course of the retrieval effort, however, as the attorney learns more about the case, about the subject matter pertinent to the case, and about the characteristics of the documents themselves, the attorney will be capable of defining relevance with greater specificity and may find it necessary to modify guidance already given. Such considerations are, of course, not unique to the e-discovery context (see, for example, Kuhlthau's excellent survey of information seeking models in [2]). In

order to be effective in these conditions, a document-retrieval method must be capable of responding to the deepening and sometimes changing conception of relevance held by the governing authority.

Accurately modeling these conditions represents a challenge for those who would evaluate the effectiveness of IR systems applied to the task. In the remainder of this paper, we turn to the challenge of evaluating the effectiveness of IR systems when applied in conditions like those that obtain in a review for responsiveness.

### III. ON EVALUATING THE EFFECTIVENESS OF SYSTEMS IN MEETING THE NEEDS OF RESPONSIVE REVIEW

In 1992, the Text Retrieval Conference (“TREC”) got underway with the first of its annual series of studies of the effectiveness of information-retrieval systems. In 2006, TREC initiated the Legal Track, with the stated mission of assessing “the ability of information retrieval technology to meet the needs of the legal community for tools to help with retrieval of business records” and, more specifically, for developing and applying “objective criteria for comparing methods for searching large heterogeneous collections using topics that approximate how real lawyers would go about propounding discovery in civil litigation” [3]. Much of the research in TREC is system-oriented, with queries held constant (for experimental control) and the retrieval effectiveness of alternate systems compared. In 2008, the Legal Track included an “Interactive Task” that was intended to model the objectives and conditions of the retrieval of documents in response to a request for production with higher fidelity. Since interaction with the single authority for relevance to clarify their intent is central to the process of responsive review, this resulted in a user-centered evaluation design.

Specifically, the task had to be designed in such a way as to capture the crucial governing role played by the senior attorney charged with overseeing a responsive review and certifying its results. To that end, the task incorporated four key design elements: (i) the introduction of a “Topic Authority” role; (ii) the provision for participants to interact with the Topic Authority; (iii) the specification that the task objective was the achievement of both high recall and high precision; and (iv) the provision that evaluation assessments were subject to final adjudication by the Topic Authority. For a complete introduction to the task, see the Guidelines for the 2008 Interactive Task [4].

#### A. *The Topic Authority*

A key element in the Interactive Task is the role of the Topic Authority (“TA”). The TA’s role, like that of the senior litigator overseeing a responsive review, is to form a conception of what is and is not responsive to a given document request (or, in the language of the task, “topic”), a conception that will derive both from considerations of genuine relevance and from considerations of the pragmatic circumstances that prevail in a given lawsuit. The role of a team that participates in the task, modeled on that of a provider of retrieval services who has

been engaged to support a review effort, is therefore to replicate, as best they can, the TA’s conception of responsiveness across the target document collection.

More specifically, the TA performs three key functions in the Interactive Task. First, the TA is a resource for teams seeking clarification as to what is and is not considered responsive for purposes of the exercise. Second, the TA provides oversight and guidance to the manual reviewers who are charged with assessing the samples of documents that will serve as the basis for measuring the teams’ performance in the exercise. Third, the TA, in the appeal and adjudication mechanism described more fully below (in Section III-D), has responsibility for making a final call on any first-pass assessment that a team has appealed.

#### B. *Interaction with the Topic Authority*

If the objective of a participating team is to replicate the TA’s conception of relevance across the test collection, provision must be made for teams to interact with the TA as a means of gaining a better understanding of what the TA considers relevant to a target topic. In the 2008 Interactive Task, this provision took the following form. Each participating team was permitted to call upon up to 10 hours of a TA’s time for each topic. The mode of interaction was largely unconstrained; teams could seek clarifications by email, arrange to speak with the TA by telephone, submit example documents for the TA to review, and so on. TAs were instructed to be free in sharing with teams any information they believed would be helpful to the performance of the task (as, in a real-world scenario, an attorney would be with a provider of retrieval services). In 2008, there was one TA per topic regardless of the number of teams working on that topic. The one restriction on the sharing of information was therefore on sharing information that had been developed exclusively through interaction with a different team; the task is designed, in part, as a test of a team’s ability to elicit from the TA a clear definition of relevance, so the restriction is necessary to prevent one team’s unrealistically benefiting from another team’s work.

#### C. *Task Objectives*

As noted above, when a party is served with a document request, it is under an obligation to make, commensurate with a reasonable good-faith effort, a complete and accurate production of documents responsive to the request. The objective, then, in the real world, is a high-recall and high-precision result, and this is the objective specified in the Interactive Task as well. Teams were asked to make a binary assessment (responsive, non-responsive) of every document in the test population; they could do this in any way that they wished (e.g., manually constructed rule-based classification, one-pass supervised machine learning, or multi-pass active learning). Their effectiveness in performing the task is measured by full-collection recall (the fraction of relevant documents that are actually produced by a system), precision (the fraction of documents produced by a system that are actually relevant), and, as a single summary metric,  $F_1$  (the harmonic mean of recall and precision).

#### D. Assessment, Appeal, & Adjudication

Values for recall, precision, and  $F_1$  are estimates based upon topic-specific samples of documents that are drawn from the full population and reviewed for relevance to the target topic. Given that the teams' objective is to replicate the TA's conception of relevance, it is obviously of crucial importance that the sample assessments, on the basis of which performance metrics are estimated, accurately reflect the TA's particular conception of relevance. To ensure that they do, the Interactive Task follows a two-step assessment procedure, a procedure whereby, in step one, volunteer assessors, under the guidance of the TA, make a first-pass assessment of the relevance of the sampled documents. Because more documents need to be assessed for a topic than can reasonably be assigned to any single volunteer, the documents to be assessed for a topic are partitioned and each partition is assigned to a different assessor. This introduces some risk of inconsistency, so in step two teams have the opportunity to appeal any first-pass assessments they believe were made in error and to have the TA render a final judgment on those appealed assessments.

Even allowing for the possibility of subsequent correction on appeal, it is in the interest of the efficiency and fairness of the task that the first-pass assessments be as accurate as possible. To that end, the task makes the following provisions for the conduct of the first-pass review. First, all assessors are provided with topic-specific guidelines, prepared by the TA, that specify the criteria by which they are to make their relevance determinations; these guidelines are essentially compilations of all the relevance guidance that the TA has given the various teams over the course of the exercise. Second, the assessors are encouraged, when they encounter a document the relevance of which remains indeterminate on the basis of the criteria provided, to ask the TA for further clarification; any such clarifications are then communicated to all assessors assigned to the topic.

While the guidelines and the opportunity for further clarification can be expected to reduce the scope for assessment error, they should not be expected to eliminate it altogether. In recognition of the fact that some assessment error will likely remain at the conclusion of the first-pass review, the Interactive Task makes provision for an appeal and adjudication mechanism as a second corrective on sample assessments. Under this provision, teams, at the conclusion of the first-pass review, are given access to all sample assessments so far entered. After reviewing the assessments, teams are invited to appeal any assessments they believe are inconsistent with guidance they received from the TA during the topic-clarification phase of the exercise. The TA then renders a final judgment on all appealed assessments. The TA is the final arbiter and there is no second round of appeal. These final, post-adjudication, assessments are the basis for estimating all task metrics.

#### IV. LESSONS FROM THE INTERACTIVE TASK IN THE TREC-2008 LEGAL TRACK

After a period of public discussion and comment by members of the Legal Track research community, the 2008

Interactive Task officially got under way on June 22, 2008. In this section, we briefly summarize some of the results of the task. We begin with a review of some specific parameters that defined the 2008 exercise, then turn to some of the key findings.

#### A. Task Specifics

Among the key elements that defined the 2008 Interactive Task were: (i) the document collection; (ii) the target topics; (iii) the Topic Authorities; and (iv) the participating teams.

1) *Document Collection*: The document collection used for the Interactive Task was the IIT Complex Document Information Processing (CDIP) Test Collection, version 1.0. This is a collection of approximately 6.9 million scanned documents that were made publicly accessible by tobacco companies under the terms of the 1998 Master Settlement Agreement. The use of scanned rather than born-digital documents introduced some additional complexity for participating teams, and OCR errors tended to depress recall values somewhat for all teams. On the other hand, by 2008 the TREC Legal Track had accumulated two years of experience with those documents, so they served as a convenient starting point for which some fully automated information retrieval systems had already been built. For a more complete description of the CDIP test collection, see the Overview of the 2006 Legal Track [5].

2) *Topics*: There were three target topics for the Interactive Task. Each took the form of a production request associated with a mock complaint (the complaint and topics are available on the Legal Track website [6]). Teams were free to submit results for one, two, or all three topics. The specific topics were as follows.

- Topic 102. Documents referring to marketing or advertising restrictions proposed for inclusion in, or actually included in, the Master Settlement Agreement ("MSA"), including, but not limited to, restrictions on advertising on billboards, stadiums, arenas, shopping malls, buses, taxis, or any other outdoor advertising.
- Topic 103. All documents which describe, refer to, report on, or mention any "in-store," "on-counter," "point of sale," or other retail marketing campaigns for cigarettes.
- Topic 104. All documents discussing or referencing payments to foreign government officials, including but not limited to expressly mentioning "bribery" and/or "pay-offs."

3) *Topic Authorities*: A single TA was assigned to each topic; the 2008 Topic Authorities were as follows.

- Topic 102. Joe Looby (of FTI Consulting).
- Topic 103. Maura Grossman (of Wachtell, Lipton, Rosen & Katz).
- Topic 104. Conor Crowley (of Daley Crowley LLP).

4) *Participating Teams*: Four teams submitted results for the Interactive Task, two from academia and two from the e-discovery industry. The teams and the topics for which each team submitted results are as follows.

- University at Buffalo. Submitted results for Topic 103.

- Clearwell Systems. Submitted results for Topics 102, 103, and 104.
- H5. Submitted results for Topic 103.
- University of Pittsburgh. Submitted results for Topics 102 and 103.

### B. Task Results

Teams took a range of different approaches to carrying out the task. In this section, we summarize some of the key findings from the 2008 running of the Interactive Task (a full description of task results can be found in the Legal Track Overview [3]).

1) *Appeal & Adjudication*: The appeal and adjudication process was used extensively for Topic 103 (which all teams completed), but much less so for Topics 102 and 104. This resulted in a substantial number of corrections to first-pass assessments for Topic 103. Aggregating all three topics, a total of 13,500 documents were sampled and assessed for evaluation purposes. Of the first-pass assessments on these documents, 966 were appealed to a TA for final adjudication. Of these 966 appeals, 762 (78.9%) were decided in favor of the appealing team (overturning the first-pass assessment); 204 (21.1%) were decided in favor of the original assessment.

The impact of the appeal and adjudication process was generally an across-the-board improvement in scores. For Topic 103, all teams saw an improvement in their  $F_1$  scores as a result of the appeals process. This did not result in a change in the relative pre- to post-adjudication rank ordering of the  $F_1$  scores by team, however.

2) *Results & Team-TA Interaction*: A fuller discussion of the recall, precision, and  $F_1$  achieved by each participant on each topic can be found in the track coordinators' Overview of the 2008 Legal Track [3]. For our purposes, what is of greatest interest is how those measures correlate with approaches to interacting with the TA. What we would like to know is whether a greater amount of interaction with the TA results in improved performance on the retrieval exercise.

In practice, teams generally made much less use of the opportunity to interact with the TA than the task permitted them to make. There was considerable variation in the amount of time used by teams; for Topic 103, for example, one team used 485 minutes of the TA's time, while another used just 5 minutes. On the whole, however, teams used only a small portion of their permitted interaction time. Setting aside the team that used 485 minutes, teams used, on average, just 60 of the 600 minutes allocated to them for interaction with the TA.

In terms of results, the teams that made only limited use of the opportunity to interact with the TAs designated for their topics (i.e., most of the teams) generally achieved fairly high precision; for the one topic in which all teams participated (Topic 103), teams that made limited use of available TA time achieved precision in the 0.70 to 0.80 range. On recall, however, these teams achieved relatively low scores; looking again at Topic 103, recall ranged from approximately 0.03 to 0.16.

We have seen that one team did make extensive use of the opportunity to interact with the TA for one topic. In terms of results, this team, like the others, achieved high precision, realizing an estimated 0.81 precision. Unlike the other teams, however, this team also achieved high recall, realizing an estimated 0.62 recall. Indeed, when documents with poor OCR are automatically culled from the collection and the scores recomputed, that team achieved in the neighborhood of 0.80 on both precision and recall.

Given the number of topics and teams that participated in the 2008 exercise, we do not have a large number of data points to work with; it does appear, however, that there is a correlation between time spent with the TA and effectiveness of the retrieval effort.

Much more research remains to be done, as we push further and ask, not simply whether interacting with the TA leads to improvements in retrieval effectiveness, but also whether there are particular methods of interacting with the authority that are more effective than others. For now, we can say that the results of the 2008 Interactive Task suggest that receptivity and responsiveness to governance by the end user is a key to effective retrieval performance in conditions like those that obtain in responsive review.

## V. CONCLUSION

We conclude with some reflections on what we have learned from the 2008 running of the Interactive Task.

We note, first, some of the strengths of the task.

- Collaboration. Bringing together information-retrieval researchers and attorneys on a real-world task fosters greater communication and collaboration in finding effective solutions to the challenges of e-discovery. For the thoughts of the individuals who filled the role of TA in the 2008 Interactive Task, see the Reflections of the Topic Authorities [7].
- Design inspiration. The relatively high fidelity of the task model can be expected to result in system and process designs that are reasonably well adapted to the target application.
- Evaluation measures. Statistically reliable estimation of precision and recall, with error bars sufficiently tight to meaningfully compare actual systems and processes were obtained.
- Practicality. The use of multiple assessors, with a process for assuring standardization, results in greater flexibility and (because of the use of volunteer labor) somewhat greater affordability that might otherwise be the case.

We also note some of the potential limitations that merit further study.

- Generalization. It is well known in IR that there is often a strong system-topic interaction, with some systems being better for some topics and other systems better for other topics. The wide range of relevance densities and a host of other factors cause this. An extensive heritage of experimentation in IR suggests that 40 topics is a practical minimum for reliable comparison of systems that

do not involve human interaction using mean effectiveness measures; human interaction induces additional variation that would likely increase this. Until a greater array of topics have been tried, those studying evaluation results will have to consider carefully the features of the specific topics used in the evaluation (in terms of complexity, nuance, yield, and so on) when interpreting the results.

- Estimation of metrics. The validity of the estimates of performance metrics is dependent on the validity of the sample assessments. As noted above, the task includes a number of quality-control measures designed to ensure that sample assessments are aligned with the target conception of relevance, chief among these the appeal and adjudication mechanism. In the 2008 exercise, however, only one topic saw extensive use of this mechanism. Going forward, we will have to investigate ways to make this mechanism more efficient (and therefore more extensively used), and we will want to explore other measures to ensure the quality of the sample assessments.
- Costs and benefits. The costs of participating in the evaluation, in terms of time and resources, can be high for both participating teams and (in the event of extensive adjudication) for the topic authority. The benefits, in terms of insights gained from these first few topics, can also be high for the research community, the vendor community, and, especially, a legal community that is seeking ways to assess how best to meet their document-retrieval challenges.
- Reusability. As is common in IR evaluation, our sampling is densest in the parts of the collection in which participating systems returned documents, and sparsest where no documents were returned by any system. As a result, the error bars will naturally be somewhat larger for our estimates of recall, precision, and  $F_1$  for subsequent use of the same relevance judgments with systems that return documents in the relatively sparsely sampled parts of the collection. The legacy of interactive experiments at TREC suggests that this may not be uncommon, since systems with human interaction have been observed to exhibit a broader range of behaviors than fully automatic systems [8]. As we build up an experience base with reuse of the collection, we will gain some ability to characterize the magnitude of this effect.
- Single Figure of Merit. Our choice of the harmonic mean of precision and recall ( $F_1$ ) is principled in that the harmonic mean is a more appropriate choice than the arithmetic mean or the geometric mean would be for ratios. But the decision to assign the same weight to recall and precision (i.e., setting  $\beta = 1$ ) was entirely arbitrary. Recall is always important in e-discovery, but in some cost-constrained settings precision can be quite important as well. Further discussion is needed to determine whether there is a setting for  $\beta$  that can garner broad agreement among the participants. Separately reporting precision and

recall allows any  $F$  measure to be computed, of course, but if we wish to make meaningful cross-system comparisons it would be useful if each participating team was optimizing for the same value of  $\beta$ .

- Scope. Conducting a review for responsiveness represents one of the most important and potentially costly retrieval tasks that may be required in the course of a lawsuit, but, as noted above, it is not the only retrieval task. Conducting a review for documents subject to a specific form of privilege, for example, represents a second important and potentially costly retrieval task to which similar methods could usefully be applied. The development of evaluation protocols, derived from the Interactive Task, that measure the effectiveness of systems at performing these additional tasks could benefit both the research and the legal communities.

We have seen that the conditions that obtain in a review for responsiveness in an e-discovery setting present unique opportunities and challenges, both for the engineering of IR systems and for the evaluation of those systems. The Interactive Task of the 2008 TREC Legal Track was designed specifically to enable the testing of IR systems in these conditions. The running of the task taught us much, both on the design side and on the execution side. We look forward to building on these lessons in another running of the task in the 2009 Legal Track.

#### ACKNOWLEDGMENT

The authors would like to thank the National Institute of Standards and Technology and TREC for continued support for our research efforts in the Legal Track; the authors would also like to thank our colleagues in the coordination of the Legal Track, Jason Baron and Stephen Tomlinson, for their always productive collaboration.

#### REFERENCES

- [1] The Sedona Conference, "The Sedona Conference best practices commentary on the use of search and information retrieval methods in e-discovery," *The Sedona Conference Journal*, no. 8, 2007.
- [2] C. C. Kuhlthau, "Inside the Search Process: Information Seeking from the User's Perspective," *Journal of the American Society for Information Science (JASIS)*, vol. 42, no. 5, pp. 361–371, Jun. 1991.
- [3] D. W. Oard, B. Hedin, S. Tomlinson, and J. R. Baron, "Overview of the TREC 2008 Legal Track," in *The Seventeenth Text REtrieval Conference (TREC 2008) Proceedings*, Nov. 2008.
- [4] J. R. Baron, B. Hedin, D. W. Oard, and S. Tomlinson. (2008) Final Interactive Task Guidelines. [Online]. Available: <http://trec-legal.umiacs.umd.edu/2008InteractiveGuidelines.pdf>
- [5] J. R. Baron, D. D. Lewis, and D. W. Oard, "Overview of the TREC 2006 Legal Track," in *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, Nov. 2006.
- [6] (2009) The Legal Track Website. [Online]. Available: <http://trec-legal.umiacs.umd.edu/>
- [7] M. R. Grossman, C. R. Crowley, and J. Looby. (2009) Reflections of the Topic Authorities. [Online]. Available: <http://trec-legal.umiacs.umd.edu/TArelections2008.doc>
- [8] E. M. Voorhees, "Variations in relevance judgments and the measurement of retrieval effectiveness," *Inf. Process. Manage.*, vol. 36, no. 5, pp. 697–716, 2000.